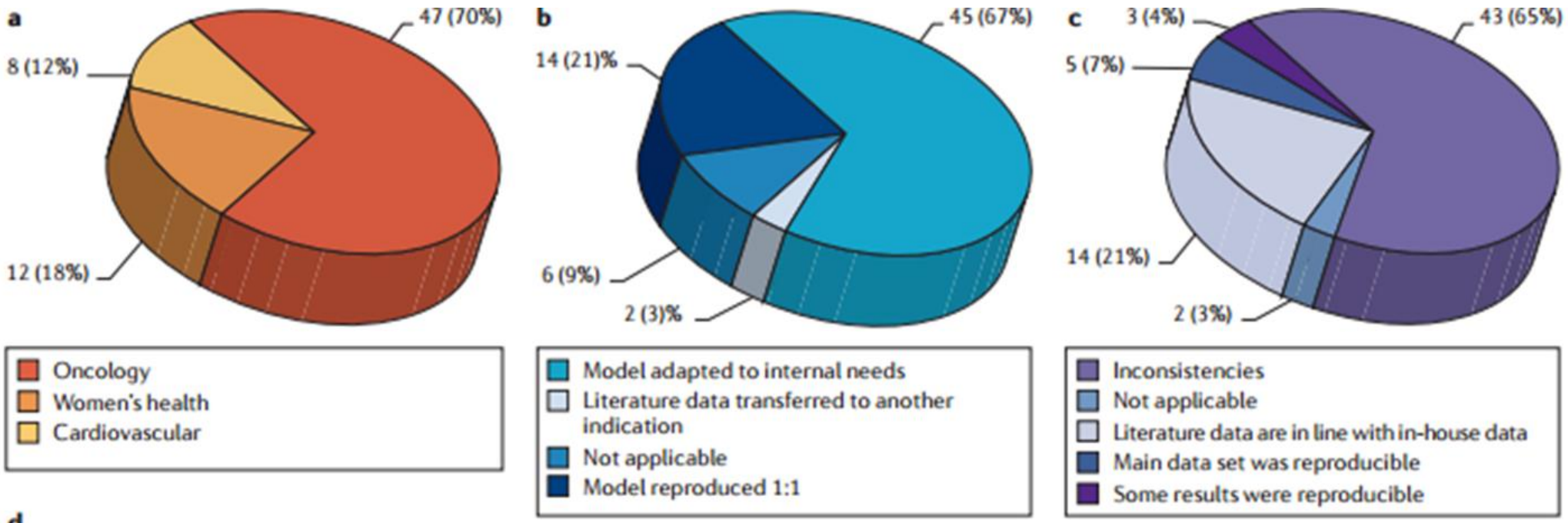
Sample Size and Power Issues

Neuroscience, Animal Models, Clinical Research

An Erosion of Faith (Academic to Industry)

- The “unspoken rule” is that at least 50% of published studies (even in top-tier journals) “can’t be repeated with the same conclusions by an industrial lab”
  - -Bruce Booth (Venture Capitalist and Atlas Venture partner)
  - “Hedging against academic risk” –Osherovich
- The success rate of Phase II clinical research has fallen from 28% to 18%
  - “Phase II failures: 2008–2010” (Arrowsmith, 2011)
  - Phase II is when effect size really comes into play

An Erosion of Faith (Academic to Industry)



d

Believe it or not: how much can we rely on published data on potential drug targets?
-Prinz et al., 2011

Power failure: Why small sample size undermines the reliability of neuroscience

Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafò

Power Failure

- Focus on the others side of the power ‘coin’ in neuroscience
  - A found significant effect represents a true effect

- Positive Predictive Value (PPV)
  - Seen in “Why most published research findings are false”
  - Not well understood (even in our own department)
  - PPV = [power x pre-study odds] / [power x (pre-study odds + type 1 error)]
    - ↓ power + ↑ type 1 error = ↓ PPV
    - ↓ power + constant type 1 error = ↓ PPV

Winner’s Curse (Proteus Phenomenon)

- The first published study is often biased towards the extreme
  - Replications find smaller or contradicting effects
- If a true effect is medium sized but researchers use small powered study, that study can only detect overestimated large effects
- “Vicious cycle” – Dr. Ken Kelley (2017)
  - Small studies produce overestimated effects which lead to small studies

Proteus

- Greek sea-god, son of Poseidon
- Protos = ‘first’
- Carl Jung defined Proteus as the personification of the unconscious
  - Powerful and shape-shifting

-Wikipedia

Winner's Curse (Proteus Phenomenon)

- Quantifying selective reporting and the Proteus Phenomenon for multiple datasets with similar bias (Pfeiffer, Bertram, & Ioannidis; 2011)
  - Analyzed a set of Alzheimer's research (1167 results from studies on 102 genetic markers)
  - Used different models to map chance of being published at different Z-values
    - Model 1: all studies have the same chance
    - Model 2: Initial study verse subsequent studies
    - Model 3: Initial study verse early replication verse subsequent publications
    - Model 4: Same as Model 3, but later publishing is dependent on first publication
      - Do contradictory results of the first publication lead to a greater chance of publication?

Winner’s Curse (Proteus Phenomenon)

- Results:
  - Initial studies have a stronger bias than subsequent studies
    - You need a greater Z-value to be published
  - Early studies tend to be biased against the result of the initial study
    - It is easier to publish contradictory results
  - HOWEVER, this was smaller than the bias initial studies have over subsequent studies

Why most discovered true associations are inflated (Ioannidis, 2008)

TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance (*P* 0.05)

True OR	Control Group Rate (%)	Sample n Per Group	Observed OR in Significant Associations	
			Median (IQR)	Median Fold Inflation
1.10	30	1000	1.23 (1.23–1.29)	1.11
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

OR = the odds an outcome will occur given an exposure divided by the odds of the outcome in the absence of the exposure

Low Power with Other Biases

- Vibration of effects: a study obtains different estimates of magnitude of effect from researcher degrees of freedom
  - Low power studies are more likely to lead to a larger vibration
  - This is especially salient in fMRI where unique programs for data analysis are the norm
    - Researchers pick different programs until one maps the data as they think it should
    - Dead Salmon study (Dr. Craig Bennet)

Low Power with Other Biases

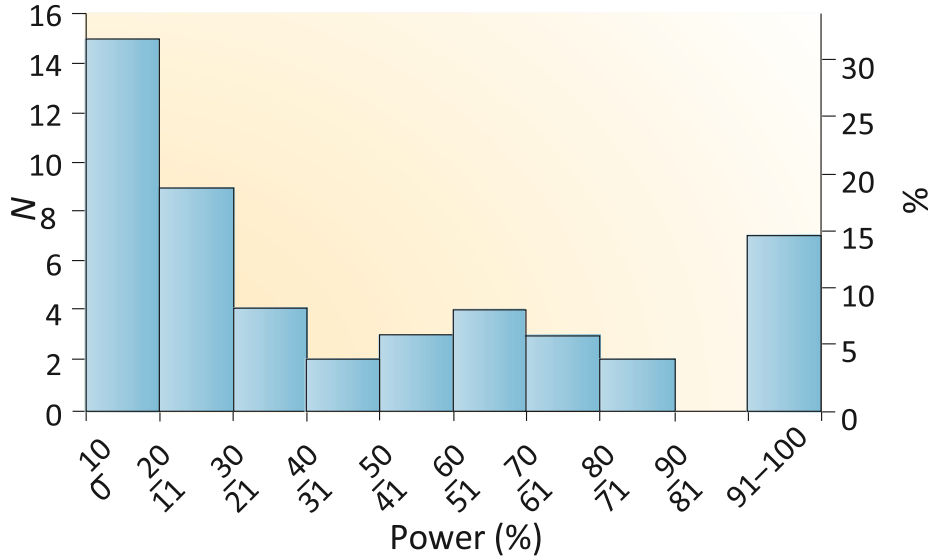
- Publication bias and selective reporting is more likely to affect small sample sizes?
  - His references only cite evidence for selective reporting (researchers are less likely to send in non-significant results of small samples)
- Worse design quality (other than sample size)?

Power of Neuroscience

- Meta-analysis of meta-analysis articles published in 2011 (included ‘neuroscience’ and ‘meta-analysis’)
  - 49 analyses with 730 primary studies
- Median power = 21%
- Test of excess statistical significance (Ioannidis & Trikalinos, 2007)
  - Basically, one uses the effect size of the meta-analysis to estimate how many studies should be significant in a given field
  - You can also compare using picked effect sizes/different error rates
  - This resulted in an estimate of 254 significant studies when 349 were found

Power of Neuroscience

- Bimodal distribution of power (many studies are less than 20% power)



- The authors argue their power calculations are ‘extremely optimistic’

Power of Neuroscience

- The authors get more specific with neuroscience fields
- Neuroimaging studies = 8% power
- “Potential reporting bias in fMRI studies of the brain” (David et al., 2013)
  - Study of correlation between sample size and significant foci reported
  - There was no correlation between sample size and significant foci reported
  - Studies with sample sizes less than 45 produce more significant foci per subject than meta-analyses

Animal Models

	Total animals used	Required *N* per study		Typical *N* per study		Detectable effect for typical *N*	
		80% power	95% power	Mean	Median	80% power	95% power
Water maze	420	134	220	22	20	$d = 1.26$	$d = 1.62$
Radial maze	514	68	112	24	20	$d = 1.20$	$d = 1.54$

One probable reason why sex effects are rarely examined/ often given contradictory results across experiments

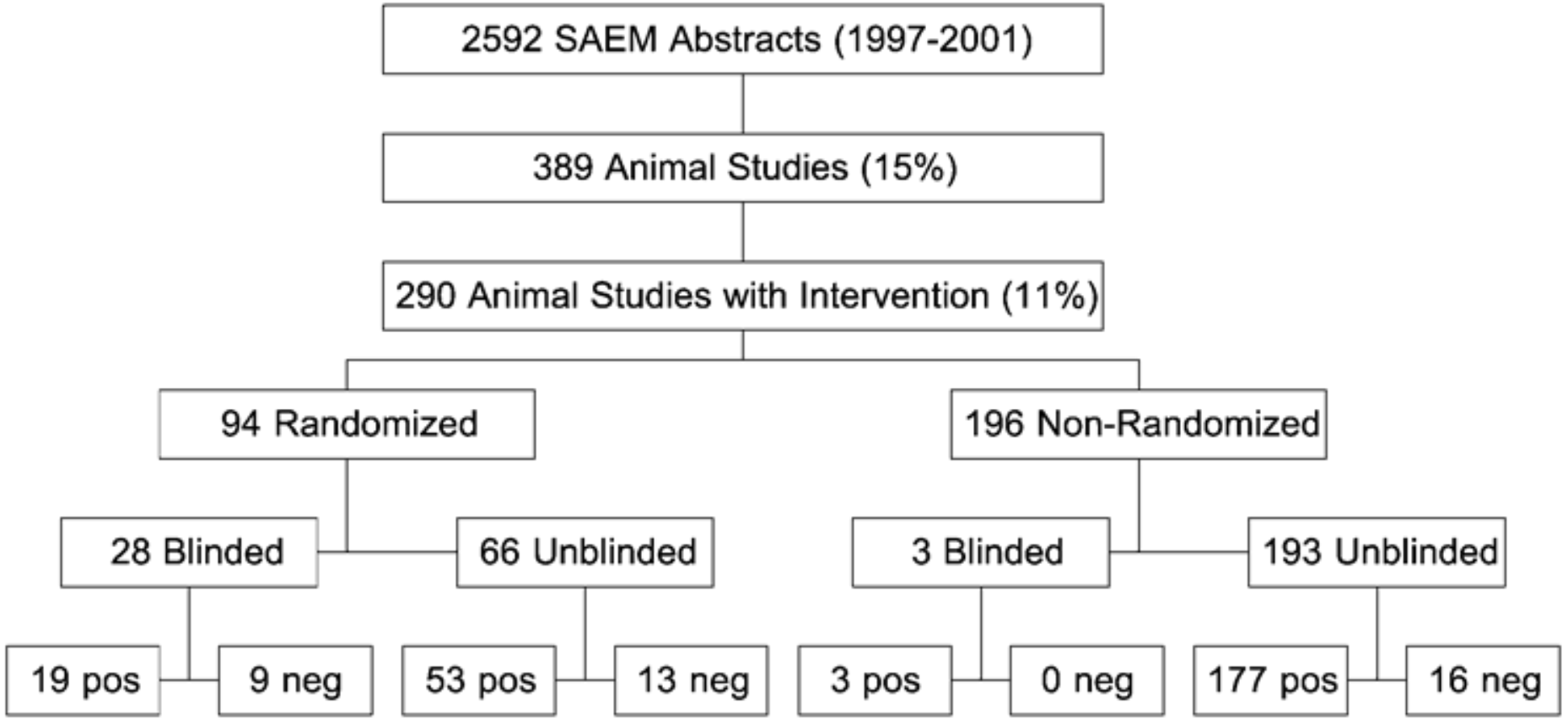
Animal Models

- Small studies give more favorable (positive) results
  - Sena et al., 2010 found 1/3 of efficacy for treatment (animal model for stroke treatment) was a result of publication bias
- Study quality is inversely related to effect size
  - Not supported by his references
  - Macleod et al., 2008 found results of animal models for stroke treatment were confounded by being non-randomized, non-blinded, ect.
  - HOWEVER, the authors ONLY mention most studies used low numbers and didn’t include it in the analysis

Animal Models

- Where is the evidence that animal research benefits humans? (Pound et al., 2004)
  - Review of studies comparing animal research to clinical research
  - Some studies matched outcomes BUT animal studies were most often irrelevant
    - Many animal studies were conducted concurrently with human studies
    - Many clinical studies went ahead when animal studies showed harm
    - The animal studies were of poor quality (not randomized or run blind)
      - Has been shown to increase chance of positive results by up to 5x
  - These results were reproduced by Perel et al., 2007 (Comparisons of treatment effects between animal experiments and clinical trials)

Randomize!



Emergency Medicine Animal Research: Does Use of Randomization and Blinding Affect the Results?
-Bebarta et al., 2003

Randomize!

- Contradicted and Initially Stronger Effects in Highly Cited Clinical Research (Ioannidis, 2003)
  - Analysis of 45 highly cited (1000+) articles in 3 clinical journals
  - 16% were later contradicted
  - 16% found larger effects than later replications
  - 24% had not been challenged
  - 5 out of 6 nonrandomized studies were later contradicted or produced larger effects than subsequently found

Ethical Implications (Animal Models)

- Animal studies need to strike a balance between using few animals but still maintaining reliable findings
- IACUC (Institutional Animal Care and Use Committee) follows the 3 Rs
  - Replace
  - Reduce
  - Refine
- The authors argue that one can waste animal life with too low of power (limiting oneself to only large effects)

Power Failure

- Conclusion
  - As neuroscientists pursue smaller and smaller effects, the sample size needs to change
    - This has not been happening
  - Low power may lead to low reproduction and poor translation from animal models to humans

Power Failure

- Steps to take
  - Increase disclosure
  - Registration of confirmatory analysis plan
  - Open data and materials
  - Incentivize replication

Current sample size conventions: Flaws, harms, and current alternatives

Bacchetti

Current sample size conventions (Bacchetti, 2010)

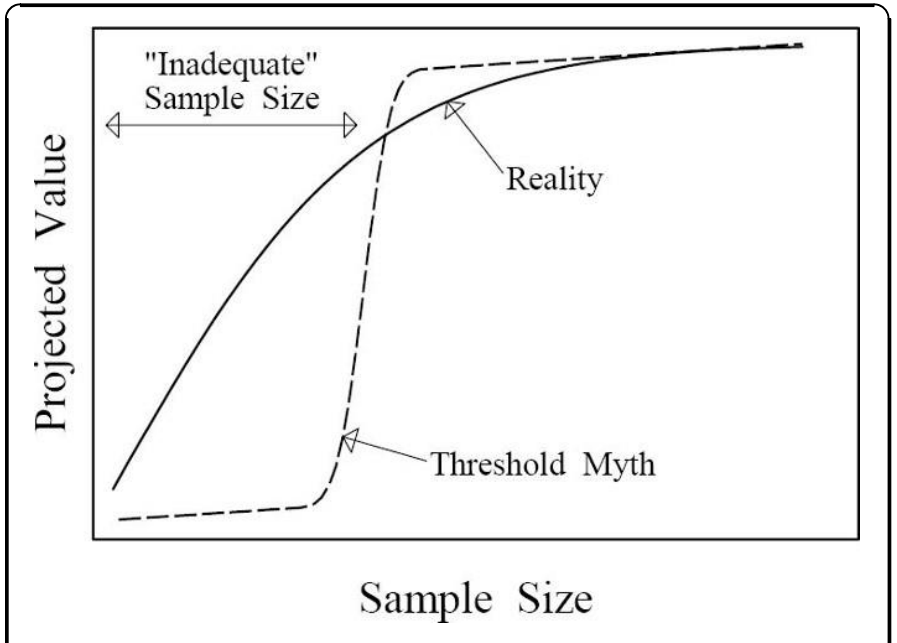
- Flaws in using 80% power as a standard
- 3 myths
  - The Threshold Myth
  - Inherent Inaccuracy
  - Design-use Mismatch
- Harms from current conventions

The Threshold Myth

- Is there a meaningful demarcation between adequate and inadequate sample sizes?
- How does projected value change with sample size
- How do you define the ‘value’ of a study?

Value terms often have square roots in their denominators*, so they follow a different pattern of projected value with sample size addition than a ‘threshold’

*Verified by his own study



Inherent Inaccuracy

- Studies with a continuous primary outcome must specify it's standard deviation (effect size?) to calculate sample size
  - This could require a lot of guess work
- Uncertainty in deciding what is important to study (effect size)

Design-Use Mismatch

- There is more to a result than the p-value
  - Effect size, confidence intervals, ect.
- Studies should be planned to optimize more than p-values
  - (A.I.P.E)

Harms from Current Conventions

- False assurance and promotion of misinterpretation
  - There is no ‘adequate’ sample size
    - 80% power and a .05 p-value cut-off still leaves room for error
  - What if your p-value is .051?
- Erosion of scientific integrity
  - Sample size game, ‘sample size samba’
    - Do our current conventions discourage truth telling about why one chooses a sample size?

Harms from Current Conventions

- Arbitrary reviewer power
  - Reviewers can use low-power to criticize unjustly
  - Does this occur?
    - Cites his own thought piece where examples are given but frequency is not proven
- Barrier to innovation
  - How are we supposed to know effect sizes of novel research?
- Wrong-way ethical standards
  - There is more to ethics than reaching 80% power

Ethics and Sample Size (Bacchetti et al., 2004)

- Focus on the Projected Net Burden of a participant
  - Participants rarely receive any direct benefit of being in the study
  - The burden per participant remains the same regardless of sample size (the total burden goes up), but the contribution of participants goes down with sample size (power)

Harms from Current Conventions

- Wasted Effort
  - We waste statisticians time by having them correct for sample size/ help interpret results
- Wasted Money
  - Participants are expensive

Alternatives

- Value of information methods
  - There are methods available that try to maximize the value of information produced while limiting the cost
  - These may be very technical/ require expertise (it's okay to 'waste' statisticians time here)
- Simple choices based on cost
  - Costs are more accurately estimated than effect size
  - Base decisions on cost per participant (cites himself)

Alternatives

- Sensitivity Analysis
  - How results change under different assumptions
  - Provides an informal assessment of the value of information that may result
    - Easier than value of information methods
- Previous similar or analogous studies
  - Use what has worked in the past

Getting There from Here

- Teach a more thorough explanation of sample size
  - Use this article!!
- Be ‘courageous’ and use different methods
- Reviewers should refrain from criticizing sample size
- Don’t report power
  - Power ‘subverts’ your study

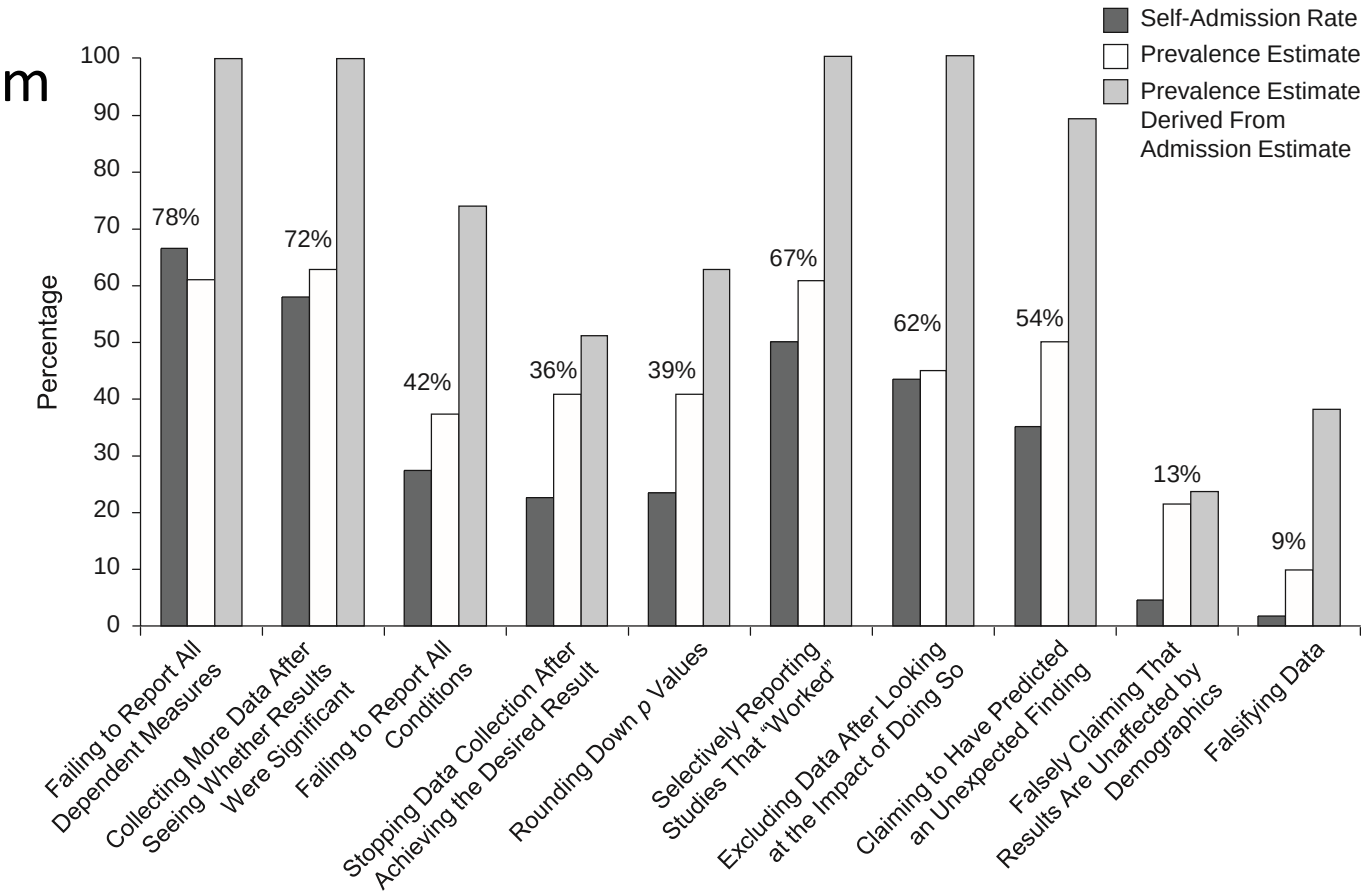
When decision heuristics and science collide

Yu, Sprenger, Thomas, & Dougherty

Decision Heuristics

- Researchers often use shortcuts in their scientific judgement and decision making
  - Researcher degrees of freedom

Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling (John, Loewenstein, & Prelec; 2012) →



Law of Small Numbers

- Tversky & Kahneman (1971): the belief that small numbers are representative of the population (leads to overconfidence)
  - Scientists are not immune?
- The authors argue this is the reason for some decisions to terminate an experiment early
  - Already achieved significance or the experiment appears like it wont ‘work’

Misconceptions Not Deceit

- The authors focus on misconceptions leading to improper research methodology
  - Not interested in purposeful deception
- “We believe that the bigger issue lies not with the small number of cases of outright fraud, but with a general lack of understanding of the relationship between a researcher’s behavior, biased sampling process, and statistical analysis”
- Are we giving too much/ too little credit to researchers?

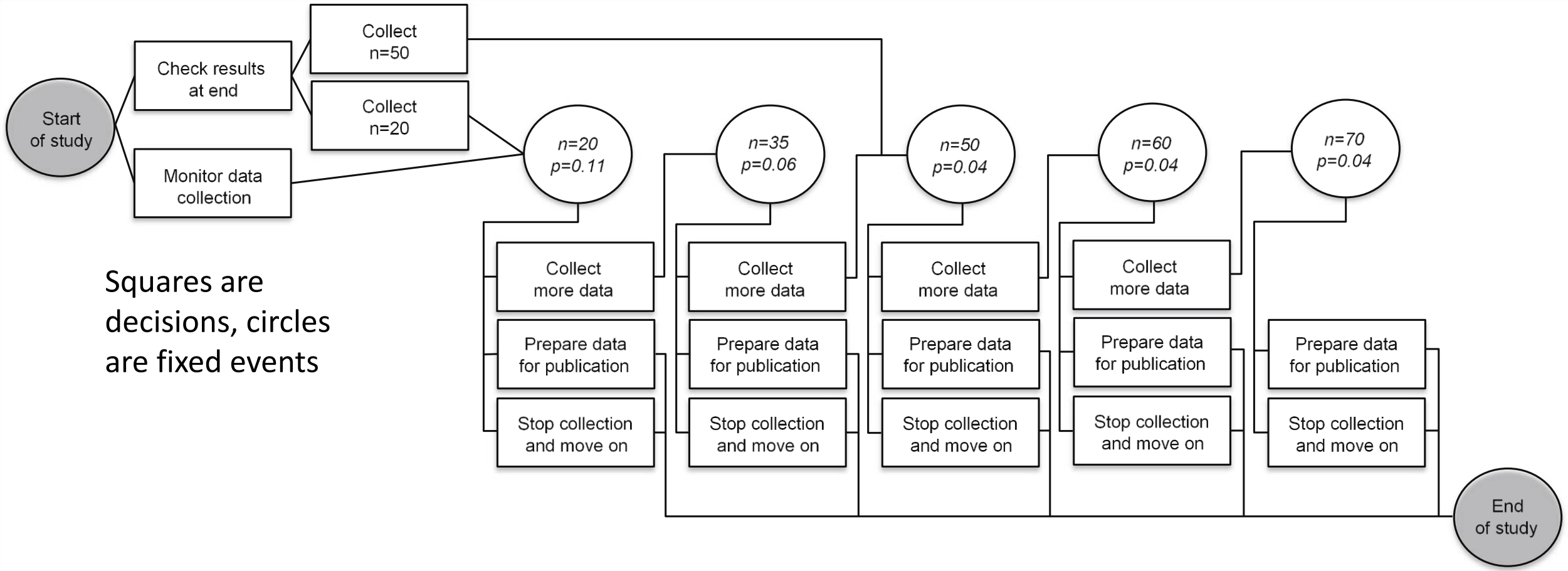
Null Hypothesis Significance Testing

- The experiment in framed in the NHST framework
- Researcher degrees of freedom are limited (theoretically)
  - Decide on a sample size a priori and stick to it
- No acting on preliminary findings
- Does the imperfect power analysis used to decide on a sample size lead to a ‘ready-made justification’ for stopping analysis early?

Methods

- 314 researchers (46.5 % worked primarily within psychology)
  - Median of 4 statistics courses
- A data collection environment was simulated where the goal was to earn promotion or tenure, managing a series of 2-independent-sample experiments while being given a limited budget
- Researchers were told the typical sample size was 20-50 per group
- Researchers were given different budgets
  - Half allowed for full-funding for purist methods, while the other half received just over half of what was needed

Methods



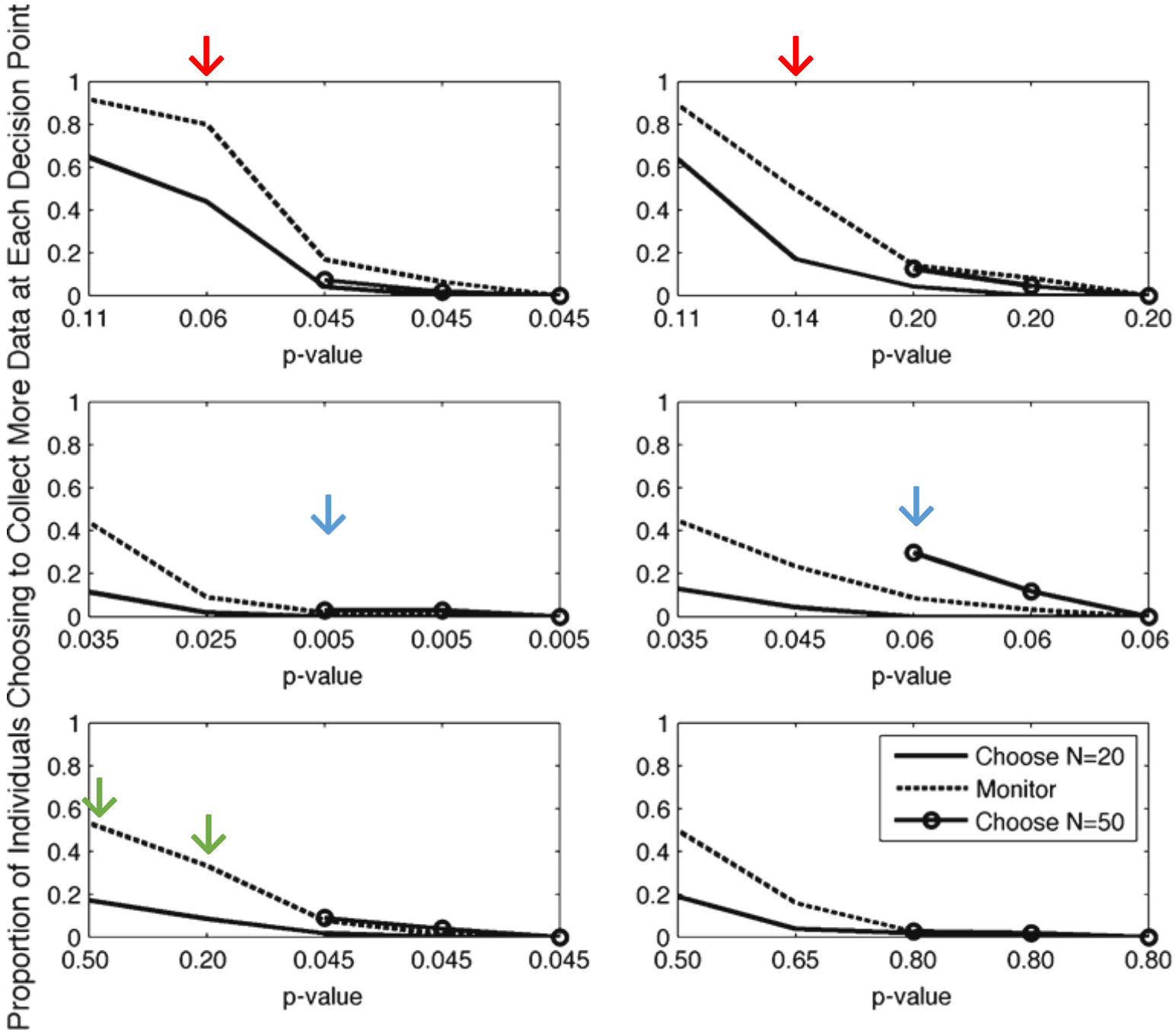
Squares are decisions, circles are fixed events

Results

- 50% of decisions were to monitor data collection
  - 32% checked after 50 subjects
  - 18% checked after 20 subjects
- Budget had no effect on data collection decisions

Fig. 2

Time 1: n= 20
Time 2: n= 35
Time 3: n= 50
Time 4: n= 60
Time 5: n= 70



Results

- The participants were also given more ‘open-ended’ questions and explicit questions about data collection practices
  - In the game = 56.7% used a p-value strategy
  - Explicit questions= 16.2% reported using a p-value strategy
  - Hypothetical questions = 49.4% would use a p-value strategy

Table 2

	Openended responses	Hypothetical choices	Game decisions, HT = 0.25	Game decisions, HT = 0.50
Purist	41.88 %	29.94 %	14.65 %	14.65 %
Low p value	9.55 %	1.91 %	21.97 %	19.43 %
High p value	6.85 %	39.81 %	12.42 %	13.69 %
Low or high p value	9.39 %	9.55 %	21.02 %	23.57 %
Other	32.32 %	18.79 %	29.94 %	28.66 %
Time	5.73 %	–	–	–
Subject availability	7.32 %	–	–	–
Money	5.10 %	–	–	–
Bayes factors	0.32 %	–	–	–
Unclassified	13.85 %	–	–	–

Simulation of Heuristics

- Stopping heuristics were simulated using Monte Carlo methodology
  - 2 group independent t-test
  - Bayes Factor were also implemented as a stopping threshold
  - Null and alternative hypotheses were used as the true hypothesis
  - Varied effect sizes (Cohen's d 0.00 – 0.80)
  - 10,000 runs for each heuristic
  - Data checking occurred every 10^{th} subject (max of 100 subjects per group)

Simulation of Heuristics

- 4 p-value strategies
  - Purist (Choose sample size a priori and stop once reached)
  - Low optional stopper (Start at n= 10 and stop if $p < .05$)
  - High optional stopper (Start at n =10 and stop is p > HT)
    - 3 variations HT= .75, .50, or .25
  - Low-high optional stopper (Start at n=10 and stop if $p < .05$ OR p > HT)
    - 3 variations HT= .75, .50, or .25
- 9 additional simulations were done using BFs as stopping heuristics
  - Start at n= 10
  - Estimate BF: If BF is greater than HT or smaller than LT stop
  - Table 3

Fig. 3

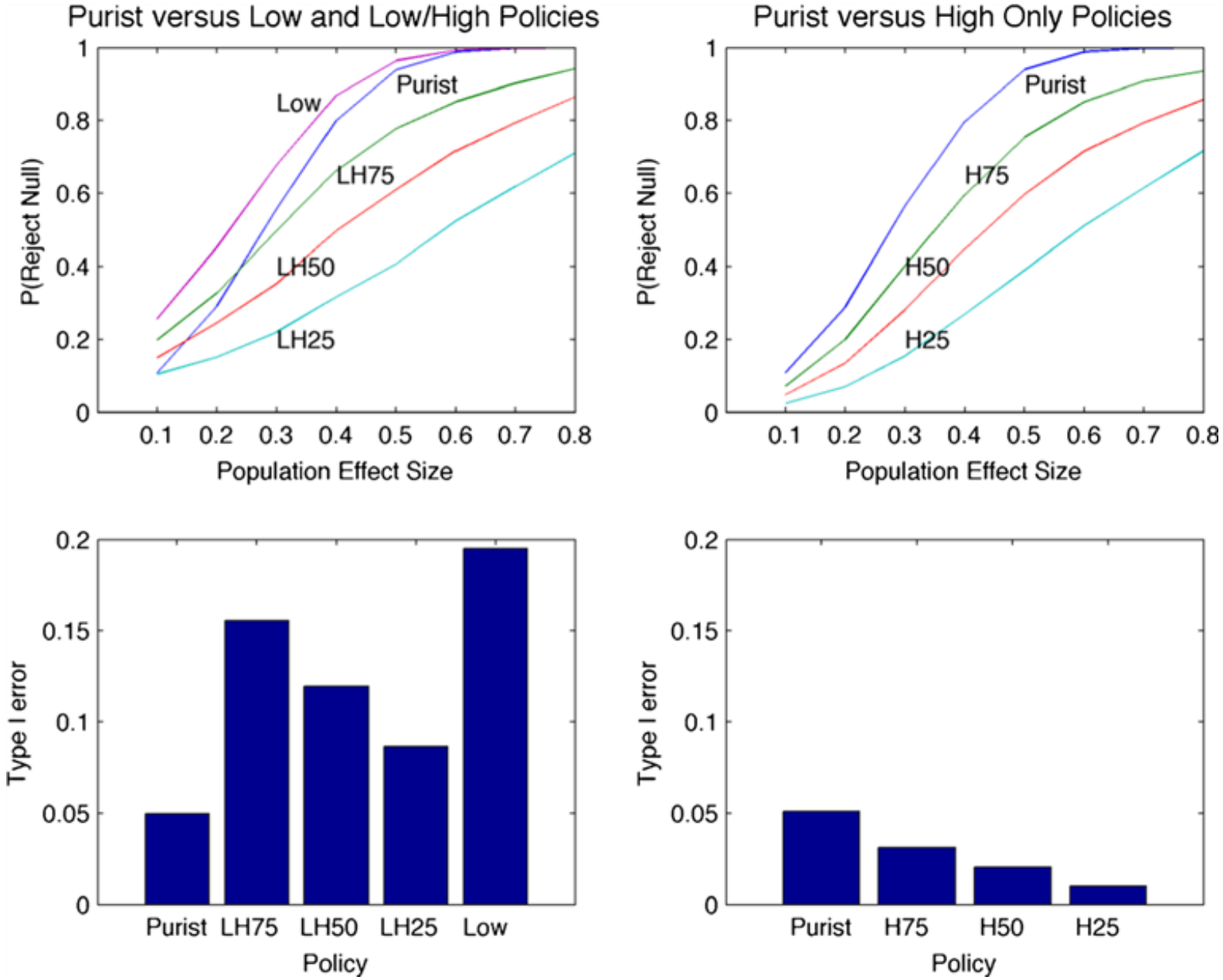

Purist versus Low and Low/High Policies
P(Reject Null)
Population Effect Size
Low
Purist
LH75
LH50
LH25
Purist versus High Only Policies
P(Reject Null)
Population Effect Size
Purist
H75
H50
H25
Type I error
Purist LH75 LH50 LH25 Low
Policy
Type I error
Purist H75 H50 H25
Policy

Fig. 4

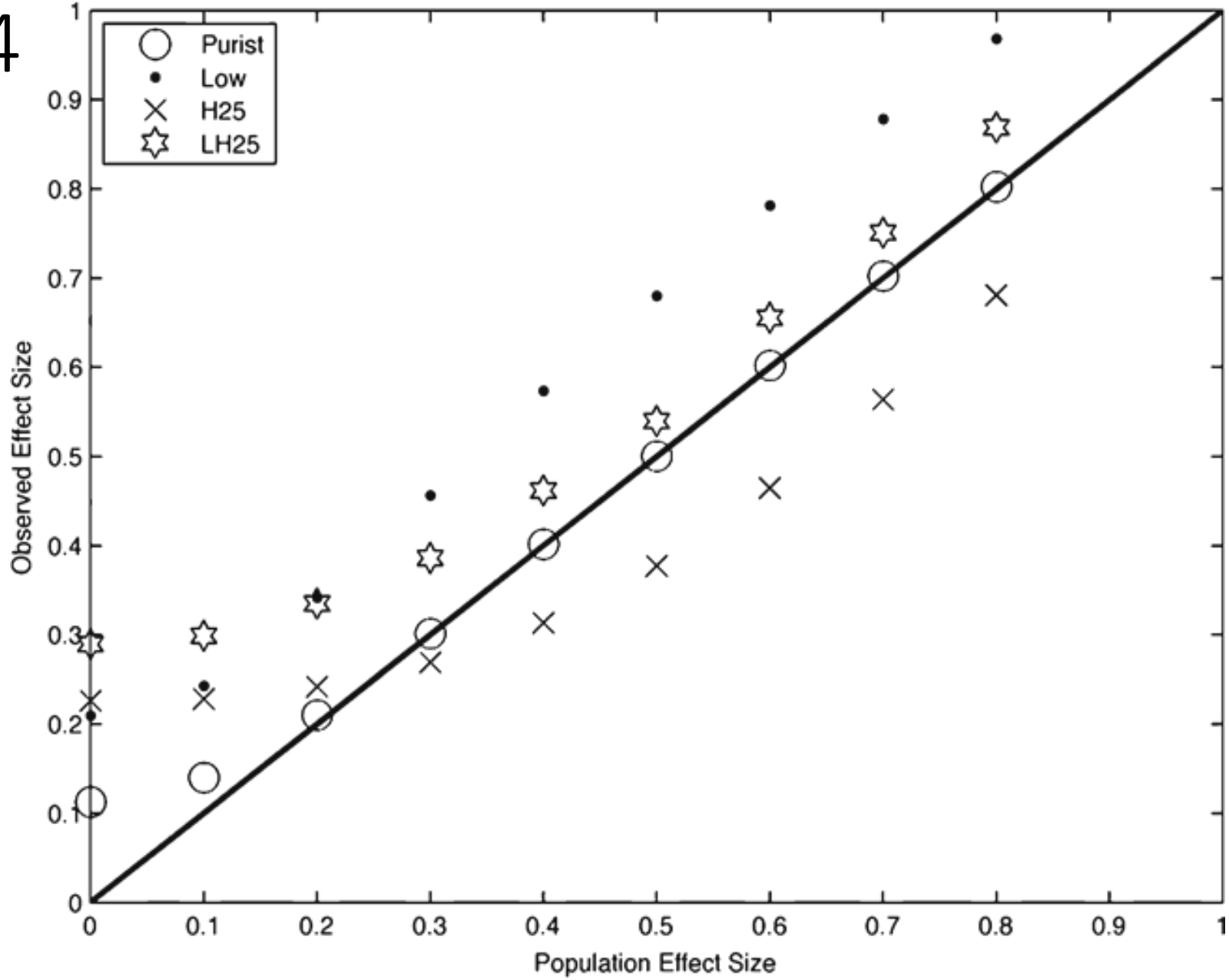
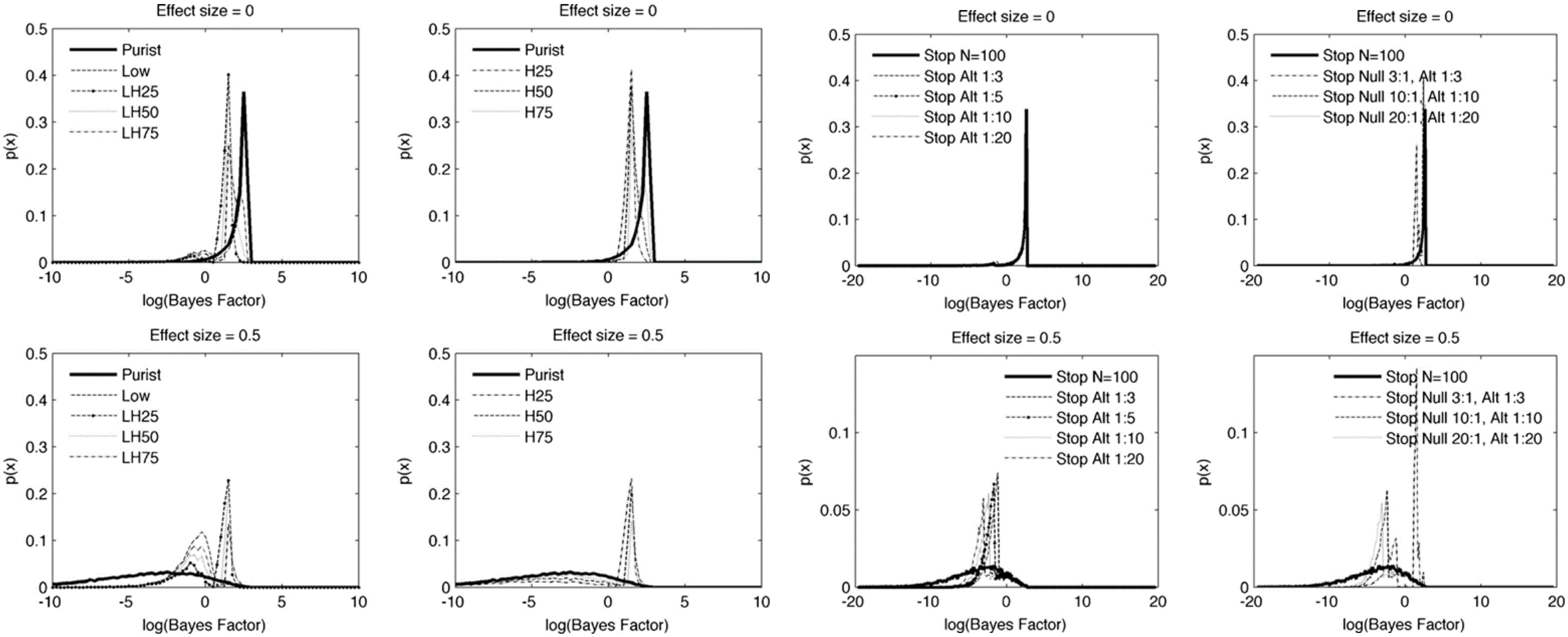
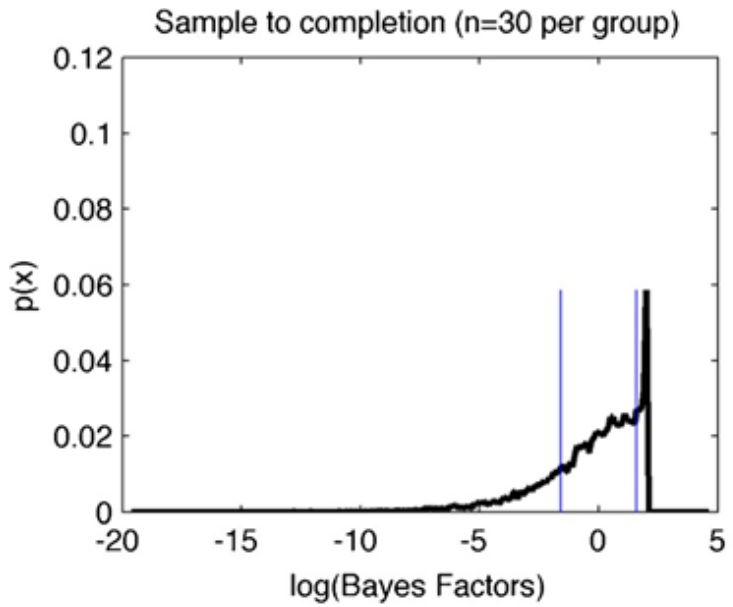

Purist
Low
H25
LH25
Observed Effect Size
Population Effect Size
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

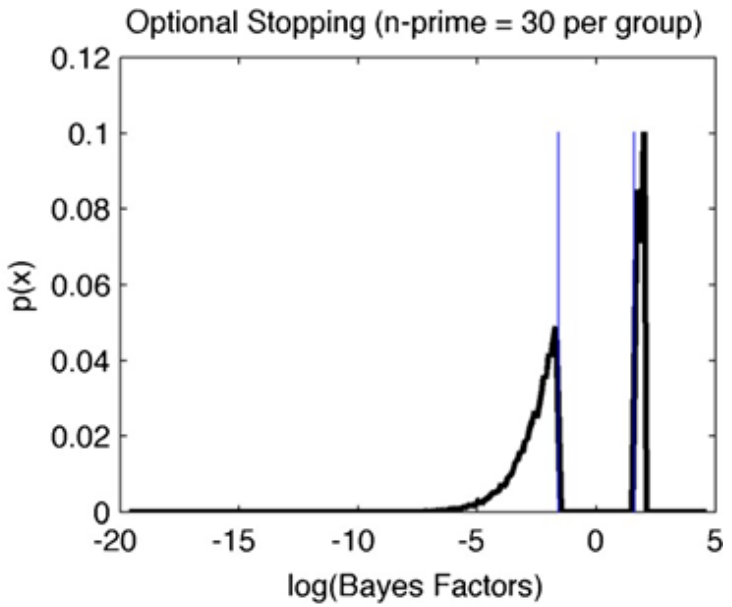
Fig. 6 & 7



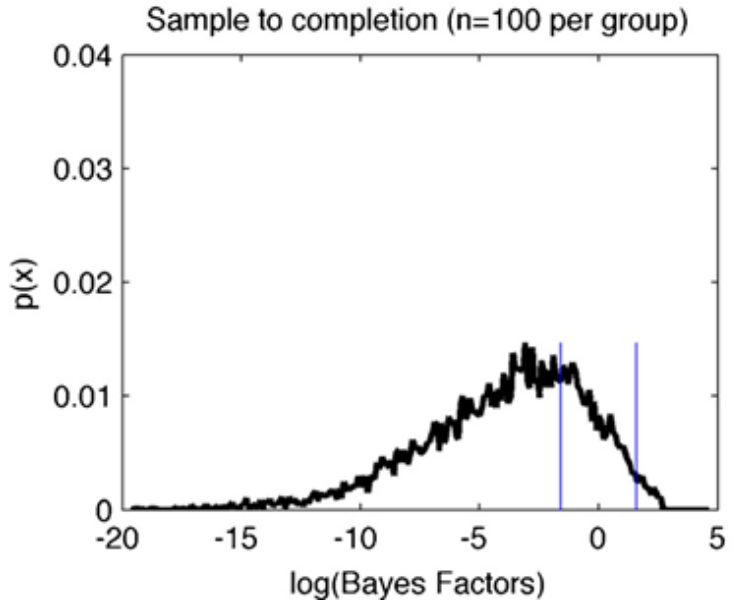
Log BF = greater than zero is evidence for null hypothesis (top line)
= less than zero is evidence for alt. hypothesis (bottom line)

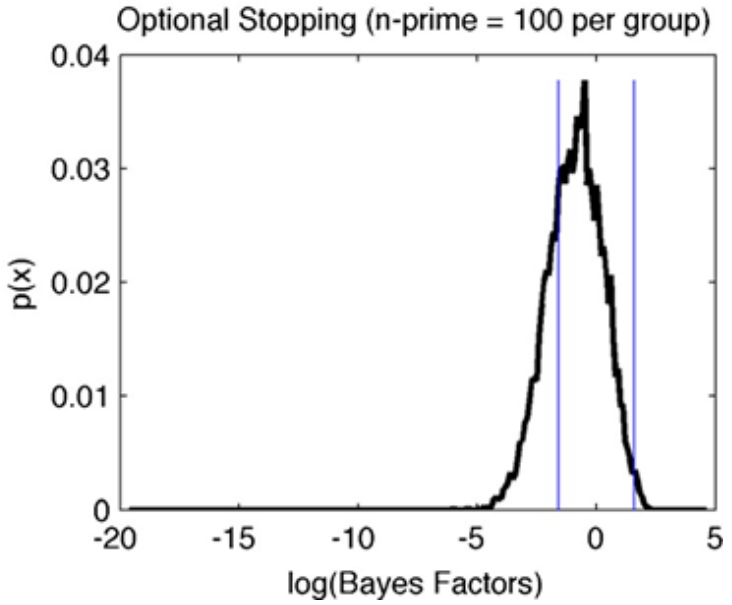
Fig. 8





Intervening causes the loss of BF values in the distribution (these are missing because one continues sampling till they are outside the range)





Planning on stopping early (but not actually stopping) changes the distribution because the distribution does not include values where you would have stopped given the chance

Limitations

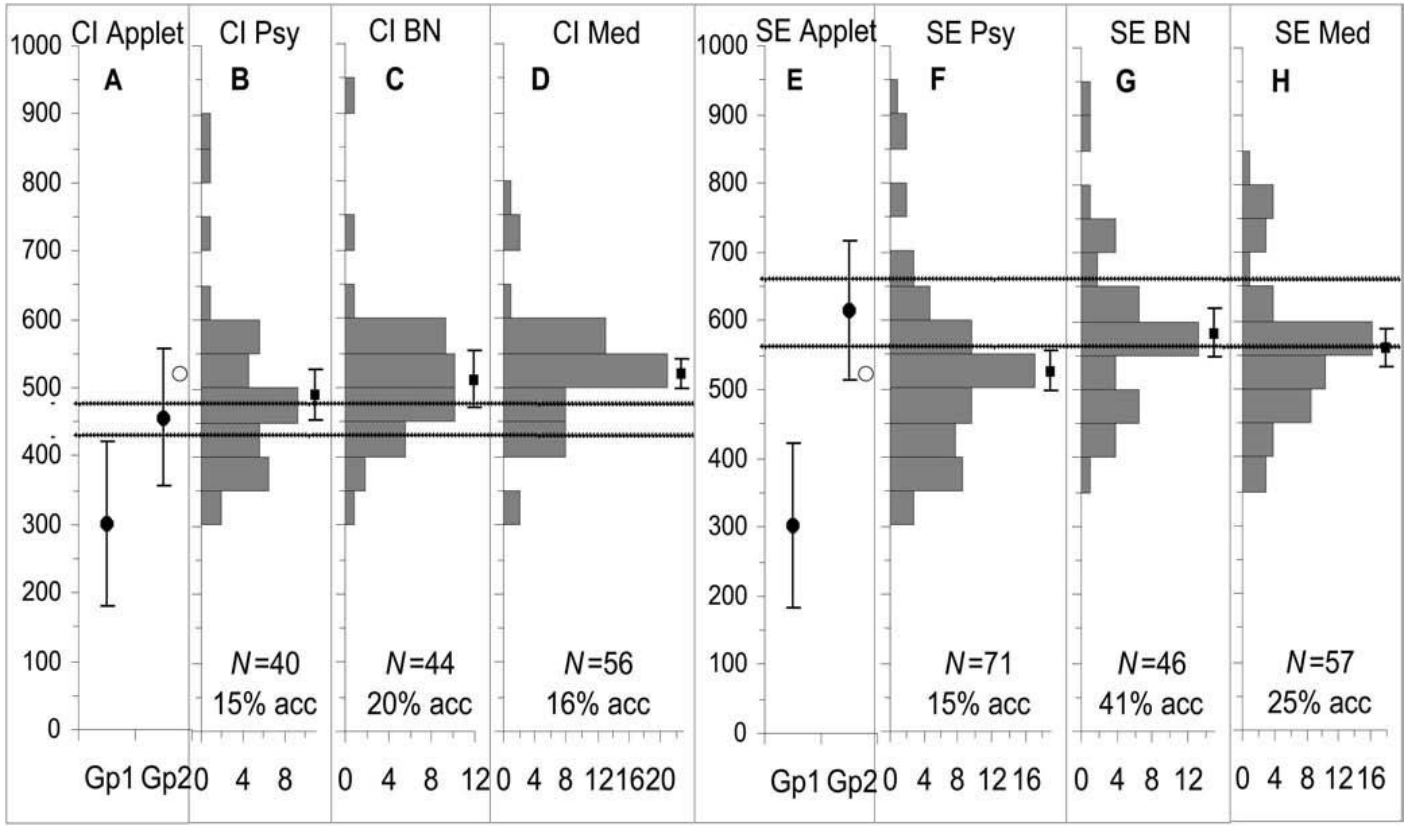
- Simulated game (different from reality?)
- Selective sample (participants responded to notice and completed experiment)

Implications

- Different heuristics lead to different patterns of
  - False positives
  - False negatives
  - Effect sizes
  - Bayes Factors (Bayesian analysis does not solve this problem)
- Scientific conclusions are influenced by these heuristics
- Heuristics will also effect meta-analyses
  - Yet another source of bias

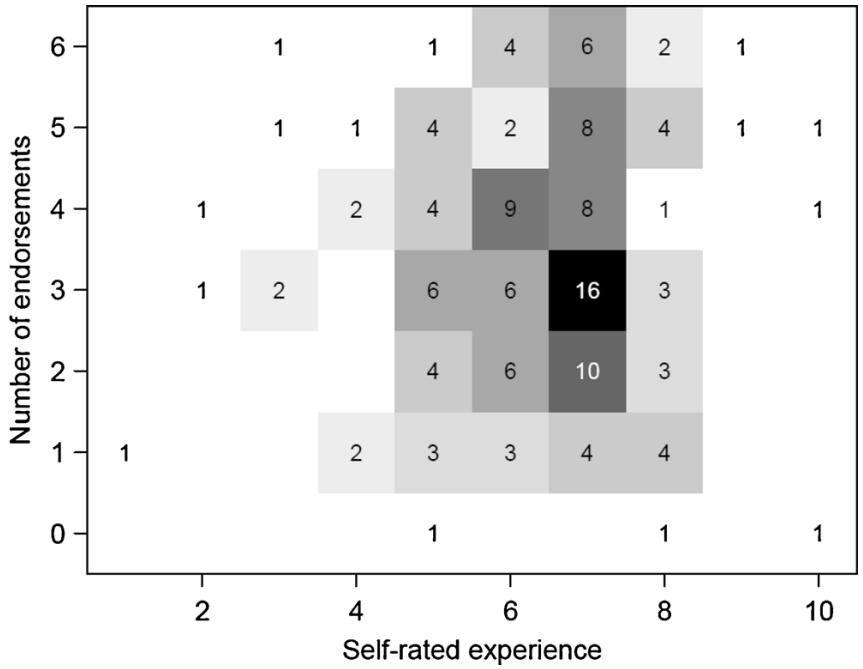
Speaking of incorrect thinking…

- Researchers misunderstand confidence intervals and standard error bars – Belia et al., 2005



Speaking of incorrect thinking…

- Robust misinterpretation of confidence intervals (Hoekstra et al., 2014)
  - Students (first year and masters) and researchers were given incorrect statements about CIs and asked to endorse which they believed were correct



Number	First-Year Students (n = 442)	Master Students (n = 34)	Researchers (n = 118)
0	2 %	0 %	3 %
1	6 %	24 %	9 %
2	14 %	18 %	14 %
3	26 %	15 %	25 %
4	30 %	12 %	22 %
5	15 %	21 %	16 %
6	7 %	12 %	11 %